**RNA3Ddesign: design sequence of RNA using tertiary structure**

Sara nourizadeh1, Bahram mohammadpour1, Parham Hafezi1, Fatemeh Zare-Mirakabad1\*

1. Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

**\* Corresponding author:** Fatemeh Zare-Mirakabad

Assitant Professor

Department of Mathematics and Computer Science

Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran

**Telephone:** +982164545674

**Email:**  [f.zare@aut.ac.ir](mailto:f.zare@aut.ac.ir)

**Abstract**

designing RNA sequence that fold to a given tertiary structure is a challenging problem in computational biology that is necessary in order to design RNAs with specific functions. Here We present RNA3Ddesign to predict the nucleotide type given the tertiary structure environment around the C5’ atom of a target base, which is repeated for every nucleotide in an RNA. Nucleotides are represented by six atoms (P, C5', O5', C4', C3', O3') on the RNA backbone in this study.

Our method consists of a four-layer deep convolutional neural network (CNN), which uses a 3D grid representation of the structure as input without extracting features manually.

We tested the performance of our method on 40 RNA with various lengths and achieved state-of-the-art performance.

**Keyword**: RNA design, RNA tertiary structure, Deep learning, 3D convolutional Neural Network, Nucleotide sequence

# **1 Introduction**

Non-coding RNAs (ncRNAs) play various essential roles at the transcriptional and post-transcriptional level[1][2][3][4]. Since the function of RNA depends on its tertiary structure, folding RNA into a specific tertiary structure is required for performing these functions. In this regard, designing RNA with a specific function, RNA inverse folding, is known as an important problem and applicable in drug design[5], gene expression control[6], RNA synthetic biology [7][8], and RNA nanostructures[9][10].

In the RNA inverse folding problem, an RNA structure is given as an input, and the goal is to find an RNA sequence folded to the target structure. Based on the type of structure, two problems for RNA inverse folding are defined RNA secondary structure inverse folding and RNA tertiary structure inverse folding.

Up to now, several methods have been developed for RNA secondary structure inverse folding, such as IncaRNAfbinv[11], MODENA[12], NUPACK[13], and others[14]. However, about RNA tertiary structure inverse folding, RNA 3D design, there is just one online web server[15] to design RNA using the tertiary structure that works for RNAs with 100nts or less. This server needs hours or days to predict ncRNAs with more than twenty nucleotides.

Based on our knowledge, there is no approach for predicting an RNA sequence with more than a hundred nucleotides based on tertiary structure.

Recently, deep-learning neural networks have achieved great success in a wide range of fields such as protein design[16], image recognition[17], and natural language processing[18].

In the field of ncRNA problems, there are some deep neural networks for RNA secondary prediction[19], RNA tertiary structure prediction[20], and RNA tertiary structure evaluation[21]. These studies led us to find a new approach for RNA tertiary structure inverse folding using deep learning.

In this paper, we propose a method, named RNA3DCNN, based on 3D convolutional neural networks (3D CNN) to predict RNA sequences with lengths of more than 100 nucleotides at a reasonable time (it takes less than 10 s). In our method, the given 3D structure of RNA is broken down into some sub-structures based on each unknown nucleotide. Then, a 3D image is made for each sub-structure using some atoms of unknown nucleotides and their neighbors. Each image is fed to the model for predicting the unknown nucleotide. One by one, each nucleotide is generated to predict the RNA sequence of the 3D structure.

Our model predicts the type of nucleotides for the 33 RNAs (released recently in NDB[[1]](#footnote-1)) with 52% accuracy. It is much better than random accuracy (=25%).

In the following subsections, we elaborate on the definition of the RNA design problem, the input and output, the architecture and configurations of the 3D CNN, the training processes, the training, and the test datasets. Finally, we discuss the results of our model.

**2 Material and methods**

**2.1 RNA 3D design Problem**

The sequence of RNA is defined as follows:

(1)

where A, C, G, and U refer to adenine, guanine, cytosine, and uracil nucleotides, respectively.

The 3D structure of RNA is noted as bellow:

(2)

where represents the coordination of atom j in nucleotide i.

In the RNA 3D design problem, the 3D structure of RNA , , is given as input. Here, we use six backbone atoms (O5’, C5’, C4', C3’, O3’, P). As an output, RNA sequence is predicted corresponding to the input structure.

**2.2 Proposed model for RNA design: RNA3Ddesign**

In this part, we first describe how to make input for RNA3Ddesign. Then, we explain the network architecture of RNA3Ddesign.

**2.2.1 Input**

We prepare the input for feeding to our model as follows:

1. The 3D structure of RNA  is given as an input, where six shows backbone atoms, O5’, C5’, C4', C3’, O3’ and P.
2. For each position , cluster contains the coordinates of six atoms of the target nucleotide and six atoms of each of fourteen closest neighbor bases ranked by the O5'-O5' distance, where .
3. All coordinates in cluster are translated and orientated to the standard coordinates by locating atoms O5’, C4’, and C3’ of the target base at the origin, x-axis, and z = 0 plane, respectively.
4. The coordinates in cluster are set into a 32\*32\*32 grid box called with each voxel being unit size (1Å × 1Å × 1Å). All voxels in the grid are zero except the voxels corresponding to the coordinates of the atoms in cluster (Figure 1).

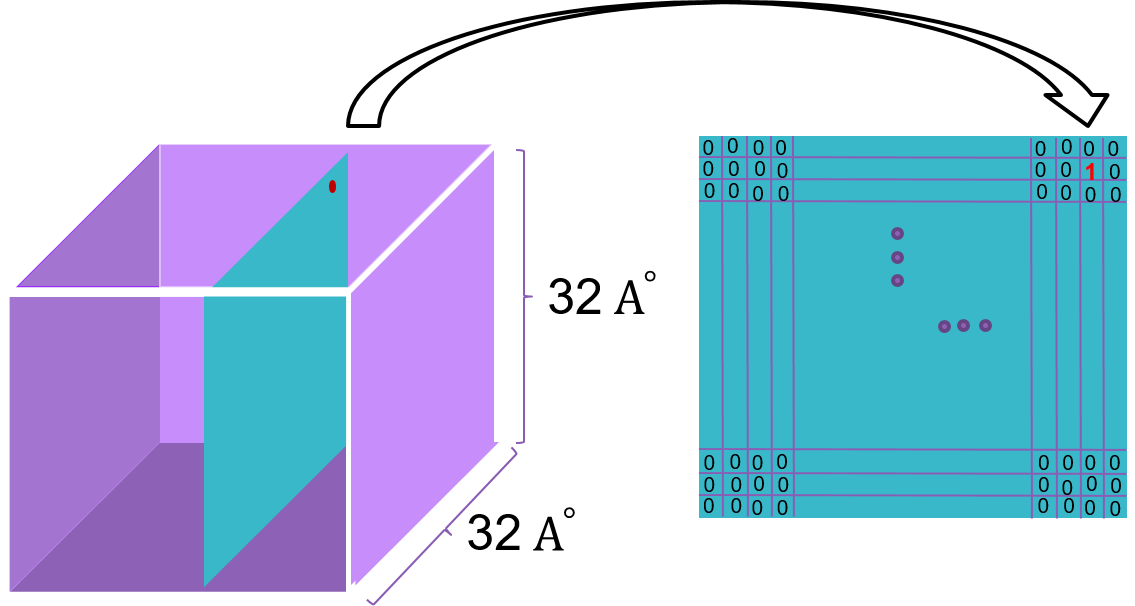


Figure 1: 3D cube representing atom distribution for input to RNA3Ddesign

**2.2.2 Network architecture**

In the past few years, the applications of the 3D convolutional neural networks have been studied in macromolecular structures, such as Protein design[25], Protein 3D structure assessment[26], RNA 3D structure assessment[21]. These methods are better than traditional machine learning approaches because they avoid manual feature extraction and engineering[??].Therefore, we use a three-dimensional convolutional network to design RNA**.**

As shown in Figure 1, we construct a 3D CNN named RNA3Ddesign described in detail as follows:

1. The input layer is a cube box with dimensions 32×32×32 (see the "Input" section).
2. Four 3D convolution layers (Conv3D) with rectified linear unit (ReLU) activation functions reduce the effect of gradient vanishing. The filters in the first two convolutional layers and the last two convolutional layers are 2 × 2 × 2 voxels and 3 × 3 × 3 voxels, respectively. The number of filters in all convolutional layers is 32. The 3D convolutional layers are used to extract 3D biochemical properties at different spatial scales.
3. The outputs of the last conv3D feed into a 2×2×2 max-pooling unit to get an output of dimension 13×13×13×60. The max-pooling layers reduce input dimensions and computational costs.
4. A fully connected layer with ReLU activation function is used to integrate information.
5. A dropout layer with a drop rate of 0.2 is defined to reduce overfitting in the neural network.
6. A fully connected layer with ReLU activation function is considered in the next part of the model.
7. A dropout layer with a drop rate of 0.2 is added.
8. A 4-dimensional SoftMax activation layer for the final output displays the probability of 4 nucleotide types of the target nucleotide.

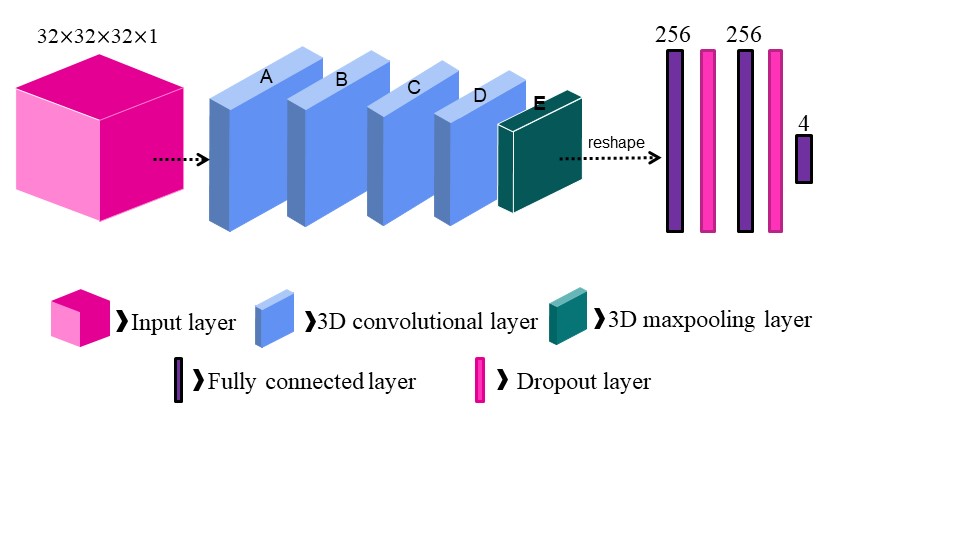


Figure 2: a graphical representation of the deep learning architecture

**2.3 Dataset**

The number of RNA sequences and structures in the Nucleic Data Bank (NDB) [22][23] has grown significantly in recent years. A list of non-redundant RNAs was downloaded with a resolution higher than 4.0 Å from the NDB website <http://ndbserver.rutgers.edu/>. Therefore, The dataset is non-redundant RNA in both sequence and structure[24]. As of May 2021, the last version of the non-redundant list of RNAs, nrlist\_3.149\_4.0A, contains 4225 RNA structures.

In addition, we pre-process data as follows:

* All RNAs with lengths less than 50nts are removed.
* Each RNA whose sequences are not consecutive is split into some consecutive RNAs.
* Pick up only the O5', C5', C4', C3', O3', and P atoms for each nucleotide of RNA.
* All RNAs with missing backbone atoms are removed.

In machine learning approaches, we require three datasets for training, validation, and evaluation. In the following, we show that how to extract these datasets from pre-processed RNAs.

The non-Redundant RNA list was further reduced to 2528 chains after pre-processing. We randomly picked 90% RNA chains (200 RNA chains) as the training dataset (TRdata) and the rest 10% (100 RNA chains) as the validation dataset (VALdata).

To evaluate our model, we downloaded “nrlist\_3.178\_4.0A” from NDB. We found 33 RNA chains in “nrlist\_3.178\_4.0A,” which are not in nrlist\_3.149\_4.0A. After pre-processing, 82 RNA chains remain. We separate these chains into continuous chains to make a test dataset named TSdata.

**استاد تا همین جا بخوانید لطفا! قسمت ریزالت نیاز داره کامل تر بشه.**

**3 Results**

In this part, we first describe the Implementation Details. Then, we explain accuracy of our model and finally result of predicting sequence of one RNA as a case study is explained.

### **3.1 Implementation Details**

Our resulting networks are summarized in Table 3. The deep 3D convolutional neural network begins with a 3D convolutional layer, followed by three sequential alternating 3D convolutional and 3D max-pooling layers, continued with two fully-connected layers, and ends with a SoftMax classifier layer.

Table 3D CNN network architecture

|  |  |  |  |
| --- | --- | --- | --- |
| stage | 3DCNN |  |  |
| layer | size | Output volume |
| Feature extraction stage | input |  | 3232321 |
|  | 3D-CONV | 2\*2\*2 512 filters | 31313132 |
|  | 3D-CONV | 2\*2\*2 256 filters | 30303032 |
|  | 3D-CONV | 3\*3\*3 32 filters | 28282832 |
|  | 3D-CONV | 3\*3\*3 16 filters | 26 |
|  | 3D-Maxpooling Layer | Stride of 2 | 13 |
| Information integrate stage | FC Layer |  | 256 neurons |
|  | Dropout  (p = 0.2) |  |  |
|  | FC Layer |  | 256 neurons |
|  | Dropout  (p = 0.2) |  |  |
| Classification stage | SoftMax Classifier |  | 4 scores |

For training, rectified linear unit (ReLU) is used as activation function except the output layer where we use SoftMax. we used categorical cross entropy as the loss function and Adam for optimization with a learning rate of 0.0001. Dropout of 0.2 was used for all fully connected layers. We set the batch size to 64. Usually, the training converges after approximately 10 epochs.

We used the Python programming language and the neural network was constructed using the Keras library[27]. Implementation detail about RAM, size of DATA , …

### **Accuracy of RNA3Ddesign**

One way to measure the accuracy of design is to estimate the sequence identity between predicted sequence and the original wild-type sequence. Accuracy of RNA3Ddesign for predicting sequence of TRdata, TSdata and evaluation dataset are shown in Table 2.

Table accuracy of RNA3Ddesign

|  |  |
| --- | --- |
| Dataset | accuracy |
| TRdata | 75% |
| TSdata | 53% |
| evaluation dataset | 50% |

We plotted confusion matrix for evaluation dataset in figure 5. Respectively, A cell (x, y) in the confusion matrix shows the number of times residue type x are predicted to be residue type y.

A

C

G

U

A C G U

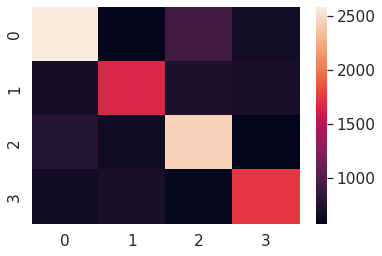


Figure confusion matrix for result of predicting evaluation dataset with RNA3Ddesign

**3.1 Case study**

In this section, we apply our proposed model for predicting the sequence of one RNA with the PDB-ID 7AQC in order to show the efficiency of our model. The predicted sequence of T chain of 7AQC using RNA3Ddesign and original sequence of this chain of RNA feed to RNAComposer[28] to predict the related tertiary structure of each sequence. Then for comparing these resulting tertiary structures, we use RNA-align[29] as its result shown in figure 6. The TM-score of this alignment is 0.57, which indicates that these two structures belong to the same Rfam[30] family as a result of being more than 0.45.

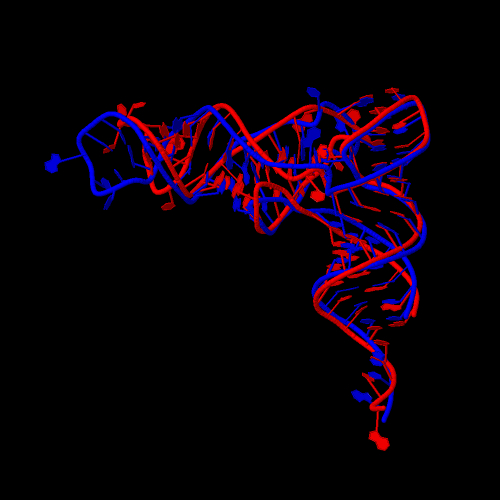


Figure result of Aligning predicted tertiary structures of the original sequence and predicted sequence with RNA3Ddesign for T chain of 7AQC

**4 conclusion**

In this study, we proposed a 3D convolutional neural network model, called RNA3Ddesign, for RNA design problem (RDP) - predicting the sequence of an RNA based on its tertiary structure – that allows us to design RNAs with a length of more than 100 nucleotides as the first study.

To improve RNA3Ddesign we can be expanding them to include more input channels, featuring more complex network architecture, and involving larger training datasets.

(چه مطالب دیگه یی باید توی این قسمت گفته بشه؟ توی مقاله های دیگه علاوه بر مطالب گفته شده اشاره میشه به مقایسه با کارهای دیگه یا برتری مدل!من مقایسه یی نداشتم)

**References:**

[1] T. Mercer, M. Dinger, J.M.-N. reviews genetics, undefined 2009, Long non-coding RNAs: insights into functions, Nature.Com. (n.d.). https://www.nature.com/articles/nrg2521 (accessed July 3, 2021).

[2] L. Srijyothi, S. Ponne, T. Prathama, C. Ashok, S. Baluchamy, Roles of Non-Coding RNAs in Transcriptional Regulation, in: Transcr. Post-Transcriptional Regul., 2018. https://doi.org/10.5772/intechopen.76125.

[3] K.C. Wang, H.Y. Chang, Molecular Mechanisms of Long Noncoding RNAs, Mol. Cell. (2011). https://doi.org/10.1016/j.molcel.2011.08.018.

[4] O. Karlsson, A.A. Baccarelli, Environmental Health and Long Non-coding RNAs, Curr. Environ. Heal. Reports. (2016). https://doi.org/10.1007/s40572-016-0092-1.

[5] I.M. Lagoja, P. Herdewijn, Use of RNA in drug design, Expert Opin. Drug Discov. (2007). https://doi.org/10.1517/17460441.2.6.889.

[6] M. McKeague, R.S. Wong, C.D. Smolke, Opportunities in the design and application of RNA for gene expression control, Nucleic Acids Res. (2016). https://doi.org/10.1093/nar/gkw151.

[7] J. Chappell, K.E. Watters, M.K. Takahashi, J.B. Lucks, A renaissance in RNA synthetic biology: New mechanisms, applications and tools for the future, Curr. Opin. Chem. Biol. (2015). https://doi.org/10.1016/j.cbpa.2015.05.018.

[8] F.J. Isaacs, D.J. Dwyer, J.J. Collins, RNA synthetic biology, Nat. Biotechnol. (2006). https://doi.org/10.1038/nbt1208.

[9] L. Jaeger, E. Westhof, N.B. Leontis, TectoRNA: Modular assembly units for the construction of RNA nano-objects, Nucleic Acids Res. (2001). https://doi.org/10.1093/nar/29.2.455.

[10] P. Guo, The emerging field of RNA nanotechnology, in: RNA Nanotechnol. Ther., 2013. https://doi.org/10.1201/b15152.

[11] M.D. Retwitzer, V. Reinharz, V. Reinharz, A. Churkin, Y. Ponty, J. Waldispühl, D. Barash, IncaRNAfbinv 2.0: A webserver and software with motif control for fragment-based design of RNAs, Bioinformatics. (2020). https://doi.org/10.1093/bioinformatics/btaa039.

[12] A. Taneda, MODENA: A multi-objective RNA inverse folding, Adv. Appl. Bioinforma. Chem. (2011). https://doi.org/10.2147/aabc.s14335.

[13] J.N. Zadeh, C.D. Steenberg, J.S. Bois, B.R. Wolfe, M.B. Pierce, A.R. Khan, R.M. Dirks, N.A. Pierce, NUPACK: Analysis and design of nucleic acid systems, J. Comput. Chem. (2011). https://doi.org/10.1002/jcc.21596.

[14] M. Lotfi, F. Zare-Mirakabad, S. Montaseri, RNA secondary structure prediction based on SHAPE data in helix regions, J. Theor. Biol. (2015). https://doi.org/10.1016/j.jtbi.2015.05.026.

[15] J.D. Yesselman, R. Das, RNA-Redesign: A web server for fixed-backbone 3D design of RNA, Nucleic Acids Res. 43 (2015) W498–W501. https://doi.org/10.1093/nar/gkv465.

[16] J. Wang, H. Cao, J.Z.H. Zhang, Y. Qi, Computational Protein Design with Deep Learning Neural Networks, Sci. Rep. (2018). https://doi.org/10.1038/s41598-018-24760-x.

[17] B.B. Traore, B. Kamsu-Foguem, F. Tangara, Deep convolution neural network for image recognition, Ecol. Inform. (2018). https://doi.org/10.1016/j.ecoinf.2018.10.002.

[18] A. Kulkarni, A. Shivananda, Deep Learning for NLP, in: Nat. Lang. Process. Recipes, 2019. https://doi.org/10.1007/978-1-4842-4267-4\_6.

[19] K. Sato, M. Akiyama, Y. Sakakibara, RNA secondary structure prediction using deep learning with thermodynamic integration, Nat. Commun. (2021). https://doi.org/10.1038/s41467-021-21194-4.

[20] R.J.L. Townshend, S. Eismann, A.M. Watkins, R. Rangan, M. Karelina, R. Das, R.O. Dror, Geometric deep learning of RNA structure, Science (80-. ). (2021).

[21] J. Li, W. Zhu, J. Wang, W. Li, S. Gong, J. Zhang, W. Wang, RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks, PLoS Comput. Biol. 14 (2018) 1–18. https://doi.org/10.1371/journal.pcbi.1006514.

[22] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan, B. Schneider, The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids, Biophys. J. (1992). https://doi.org/10.1016/S0006-3495(92)81649-1.

[23] B. Coimbatore Narayanan, J. Westbrook, S. Ghosh, A.I. Petrov, B. Sweeney, C.L. Zirbel, N.B. Leontis, H.M. Berman, The Nucleic Acid Database: New features and capabilities, Nucleic Acids Res. (2014). https://doi.org/10.1093/nar/gkt980.

[24] N.B. Leontis, C.L. Zirbel, Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking, in: 2012. https://doi.org/10.1007/978-3-642-25740-7\_13.

[25] Y. Zhang, Y. Chen, C. Wang, C.C. Lo, X. Liu, W. Wu, J. Zhang, ProDCoNN: Protein design using a convolutional neural network, Proteins Struct. Funct. Bioinforma. (2020). https://doi.org/10.1002/prot.25868.

[26] R. Sato, T. Ishida, Protein model accuracy estimation based on local structure quality assessment using 3D convolutional neural network, PLoS One. (2019). https://doi.org/10.1371/journal.pone.0221347.

[27] F. Chollet, Keras: The Python Deep Learning library, Keras.Io. (2015).

[28] M. Biesiada, K. Pachulska-Wieczorek, R.W. Adamiak, K.J. Purzycka, RNAComposer and RNA 3D structure prediction for nanotechnology, Methods. (2016). https://doi.org/10.1016/j.ymeth.2016.03.010.

[29] S. Gong, C. Zhang, RNA-align: Quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA, Bioinformatics. (2019). https://doi.org/10.1093/bioinformatics/btz282.

[30] I. Kalvari, E.P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, E. Rivas, S.R. Eddy, R.D. Finn, A. Bateman, A.I. Petrov, Rfam 14: Expanded coverage of metagenomic, viral and microRNA families, Nucleic Acids Res. (2021). https://doi.org/10.1093/nar/gkaa1047.

1. Nucleicacid DataBase (http://ndbserver.rutgers.edu/) [↑](#footnote-ref-1)